Journal of Vegetation Science 2: 587-602, 1991

Generalized additive models in plant ecology

Yee, Thomas W.[1] **& Mitchell, Neil D.**[2]

[1]*Department of Mathematics and Statistics, University of Auckland, Private Bag, Auckland, New Zealand; Fax +64 9 7327934; E-mail YEE@MAT.AUKUNI.AC.NZ;*
[2]*Department of Botany, University of Auckland, Private Bag, Auckland, New Zealand; E-mail ND01@CCU1.AUKUNI.AC.NZ*

Abstract. Generalized additive models (GAMs) are a nonparametric extension of generalized linear models (GLMs). They are introduced here as an exploratory tool in the analysis of species distributions with respect to climate. An important result is that the long-debated question of whether a response curve, in one dimension, is actually symmetric and bell-shaped or not, can be tested using GAMs. GAMs and GLMs are discussed and are illustrated by three examples using binary data. A grey-scale plot of one of the fits is constructed to indicate which areas on a map seem climatically suitable for that species. This is useful for species introductions. Further applications are mentioned.

Keywords: Additive model; Case-control study; Direct gradient analysis; Gaussian logit curve; Generalized linear model (GLM); Logistic regression; Plant species distribution; Plant species-climate relationship; Response curve; Smoother.

Abbreviations: GLM = Generalized linear model; GAM = Generalized additive model; GAIM = Generalized additive interactive model; BIOCLIM = Bioclimatic Prediction System.

Introduction

Generalized linear models (GLMs) as proposed by Nelder & Wedderburn (1972) have been successfully applied in ecological research (e.g. Austin & Cunningham 1981, Nicholls 1989, Austin, Nicholls & Margules 1990). This approach has enabled biologists to model species responses to a wide range of environmental data types (such as discrete, categorical, ordinal and continuous data) under a single theoretical and computational framework. The theory of GLMs has been well developed (McCullagh & Nelder 1989; for an introductory treatment see Dobson 1990). Literature describing logistic regression, a specialized form of GLM for binary data include Jongman, ter Braak & van Tongeren (1987), Hosmer & Lemeshow (1989), and van Houwelingen & le Cessie (1988). GLMs are commonly fitted by the GLIM (Numerical Algorithms Group 1987), GENSTAT (Payne et al. 1987) and S-PLUS (Becker, Chambers & Wilks 1988) software packages. Additional software with logistic regression capabilities include SAS (SAS Institute Inc. 1988), BMDP (Dixon 1987) and SYSTAT (Wilkinson 1987).

The use of GLMs in plant ecology has several drawbacks compared with the modelling approach advocated here; most importantly, GLMs are not as readily exploratory in nature. Ter Braak & Gremmen (1987) observe that in practice, the correct model to be fitted is rarely known and diagnostic plots seldom appear to be carried out to test the validity of the models (Landwehr, Pregibon & Shoemaker 1984).

Generalized linear models

Generalized additive models (GAMs) are a nonparametric extension of GLMs, so it is useful to review the main points of GLMs before describing GAMs in detail.

For GLMs we have sample data $(Y_i, \boldsymbol{x}_i)$ $(i = 1,2,.., n)$ where Y_i is the response variable, n is the sample size and $\boldsymbol{x}_i = (x_{i1}, x_{i2},.., x_{ip})^T$ is a vector of p explanatory variables or covariates. When $p = 1$, $\boldsymbol{x}_i$ may be written as x_i. The Y_i are independent and have a distribution belonging to the exponential family. This family contains many common distributions such as the normal (Gaussian), Poisson, binomial, negative binomial, geometric, beta, exponential and gamma. The mean of the response variable at $\boldsymbol{X} = \boldsymbol{x}$, namely, $\mu_i = \mu_i(\boldsymbol{x}) = E(Y_i) = E(Y_i|\boldsymbol{x})$, is related to the covariate information by

$$g(\mu) = \alpha + \beta^T \boldsymbol{x} = \alpha + \sum_{j=1}^{p} \beta_j x_j \qquad (1)$$

where g is a prespecified function called the link function, α is the intercept or constant term and $\beta = (\beta_1, \beta_2,\ldots, \beta_p)^T$ is a vector of regression coefficients. The right hand side of (1) is a plane in p-dimensional space. The purpose of the link function is to transform (link)

the mean of Y to lie on a plane in this p-dimensional space.

The deviance is used to measure goodness of fit of a model. The deviance, a generalization of the residual sum of squares in ordinary regression, is a function of the data and of the fitted values, and when divided by a scale factor is asymptotically chi-squared. To test if a smaller model (i.e. one with fewer variables) is applicable given a larger model, it is only necessary to examine the increase in the deviance and to compare it to a chi-square distribution with degrees of freedom equal to the difference in the number of parameters in the two models (as each parameter or term has 1 degree of freedom). This enables a test as to which variables can be deleted to form a smaller model or which variables need to be added to form a larger model. Smaller deviances indicate a better fit.

In this study only binary data were available; the following discussion deals with GLMs applied to data of this form.

Logistic regression

GLMs will be contrasted with GAMs in the context of logistic regression models for binary (e.g. presence/absence) data. Let $Y = 1$ or 0 denote the presence/absence of a species respectively and $p(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ be the probability the species is present in a quadrat of fixed size when $\boldsymbol{X} = \boldsymbol{x}$. In the literature $p(\boldsymbol{x})$ is referred to as the 'presence-absence response curve' (ter Braak & Looman 1986) and its use is an example of direct gradient analysis - the species' probability of presence (or abundance) is described as a function of measured environmental variables.

For binary data it is more convenient to model

$$\text{logit}\{p(\boldsymbol{x})\} = \log\{p(\boldsymbol{x})/(1-p(\boldsymbol{x}))\}$$

rather than $p(\boldsymbol{x})$ itself because of technical problems caused by the constraint that $p(\boldsymbol{x})$ lies between 0 and 1. In the context of GLMs we have $\mu(\boldsymbol{x}) = p(\boldsymbol{x})$ and $g(p) = \text{logit}(p)$. The logit transformation stretches the interval 0 to 1 to $-\infty$ to $+\infty$. If $\text{logit}\{p(\boldsymbol{x})\}$ could be plotted against the individual variables, this would define a (regression) surface. GLMs attempt to approximate this unknown regression surface using a restricted class of parametric terms. For example,

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2,$$

which consists of linear and quadratic terms in the variables. This is known as the 2 dimensional Gaussian logit model with no interaction term (see below). The one dimensional Gaussian logit model is

$$\begin{aligned}\text{logit}(p(x)) &= \alpha + \beta_1 x + \beta_2 x^2 \\ &= a - (x-u)^2/(2t^2). \end{aligned} \quad (2)$$

In the second formulation, u is often called the species' optimum or indicator value and t its tolerance (a measure of ecological amplitude). The resulting presence-absence response curve is

$$\begin{aligned} p(x) &= \exp(\alpha + \beta_1 x + \beta_2 x^2)/\{1 + \exp(\alpha + \beta_1 x + \beta_2 x^2)\} \\ &= 1/\{1 + \exp(-\alpha - \beta_1 x - \beta_2 x^2)\} \end{aligned}$$

which is symmetric and bell-shaped. This popular model is known as the 'Gaussian logit curve' (Jongman, ter Braak & van Tongeren 1987).

Drawbacks of GLMs

Sometimes GLMs are not flexible enough to approximate the true regression surface adequately (e.g. the Gaussian logit only accomodates symmetric bell-shaped curves whereas the true curve may be asymmetrical). Even by adding extra terms (e.g. a cubic term), the approximation may still be inadequate. GAMs allow a wider range of response curves to be modelled, of which GLMs become a special case.

For plant species, there has been a long debate as to the appropriate shape that response curves should have. For example, Whittaker (1956) observed that species typically show unimodal response curves. In that study, he proposed that species' response curves could approximate normal curves in the sense of being symmetric and bell-shaped. Since then, these curves have been promulgated in all areas of biological literature and have become almost a basic tenet in plant ecology (Austin 1979). However, these models have been under attack and heavily criticized. Austin & Smith (1989) have recently stated "When applying niche theory to plants … the assumption of bell-shaped response curves for species [is] unrealistic." They comment that most species responses appear skewed and that "the bell-shaped pseudo-gaussian response curves for both fundamental and realized niche responses are unrealistic for plants." It will be seen later that, for a single gradient, GAMs provide a convenient test for these two hypotheses. Even if the data were bell-shaped, a Gaussian logit model may not be able to model them correctly since $p(x)$ for a Gaussian logit model has a flatter top if $p(u)$ is approximately unity, than does a normal probability curve (ter Braak & Looman 1986).

There are many examples of data sets in the literature which are bimodal or skewed. For example, Austin (1987) found that positive-skewed curves were characteristic of major tree canopy species in Eucalypt forests in southern Australia. Another example is Austin,

Nicholls & Margules (1990), who analysed five species using GLMs and found bell-shaped, skewed and complex response curves. Even with the transformation of variables and the addition of higher order terms, GLMs can still be an inadequate modelling procedure.

When using GLMs (or any statistical procedure), care must be taken that a sensible result is obtained. For example, if (2) is fitted to data and the estimate of β_2, namely $\hat{\beta}_2$, is positive this would mean u is a minimum rather than a maximum. Such a model clearly contradicts biological reality. Hence, in the first case, the Gaussian logit curve (2) should be rejected if $\hat{\beta}_2 > 0$.

Generalized Additive Models

GAMs are data-driven rather than model-driven; that is, the resulting fitted values do not come from an a priori model. The rationale behind fitting a nonparametric model is that the structure of the data should be examined first, before fitting an a priori determined model. Although this can be done using GLMs (e.g. Austin, Nicholls & Margules 1990) by categorizing the data, it is inefficient.

GLMs relate the mean response to the $\boldsymbol{x}$ variables via

$$g(\mu) = \alpha + \beta^T \boldsymbol{x} = \alpha + \sum_{j=1}^{p} \beta_j x_j .$$

GAMs relax this to simply

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(x_j),$$

where the f_j are unspecified smooth functions. In practice the f_j are estimated from the data by using techniques developed for smoothing scatterplots. There are many types of scatterplot smoothers e.g. the running lines, running means, running medians, cubic splines, B-splines, the Lowess of Cleveland (1979) and the Supersmoother of Friedman & Stuetzle (1981). For an introduction to smoothers see Goodall (1990) and Hastie & Tibshirani (1990). In order to make the functions f_1, f_2,…, f_p unique, they are constrained to be centered about zero; that is, $E(f_j(x_j)) = 0$ for all j. If $f_j(x_j) = \beta_j x_j$ for all j, the GAM is a GLM.

Thus GAMs allow the data to determine the shape of the response curves, rather than being limited by the shapes available in a parametric class. As a result, features such as bimodality and pronounced asymmetry in the data can be easily detected. For this reason GAM modelling provides a better tool for data exploration than GLM modelling.

The regression surface, $g(\mu(\boldsymbol{x}))$, is expressed as a sum of functions of each variable, so that each explanatory variable has an additive effect. Consequently, one can interpret the contribution of each variable by examining each function. This is a crucial concept. However, as explained below, such an interpretation holds only if there are no interactions.

The value in having the data suggesting the response curve is illustrated by a recent paper by Austin, Nicholls & Margules (1990). In this paper, one of the reasons why three continuous variables (mean annual temperature, mean annual rainfall and solar radiation index) were categorized was so that the magnitudes of the model coefficients from each category would indicate the shape of each species' response. This is an attempt to do what GAMs can do much better! In detail, they treated each environmental variables as a factor by dividing each one into a number of classes: rainfall into nine classes of 200 mm, temperature into 12 classes of 1 °C and solar radiation into five classes of 0.1 units. From this they fitted

$$\text{logit } p = \alpha_i + \beta_j + \tau_k, \quad i = 1,..,9;\ j = 1,..,12;\ k = 1,..,5 \qquad (3)$$

and plotted the fitted α_i's, β_j's and τ_k's (called the coefficient diagram) to reveal the shape of the transformed response and whose shape suggested an appropriate continuous function for each variable. Clearly, GAMs do this but are more general, flexible, robust and efficient than this categorical/discrete modelling procedure. The work by Austin, Nicholls & Margules (1990) shows the application and value of a nonparametric fit in this context.

Note that the three variables in (3) do not interact. For this reason, the authors wrote "Interpretation of the coefficient diagram, to assess an appropriate polynomial function, is possible due to the demonstrated independence of the three factors." The same is true with GAMs - the contribution of each variable can be interpreted separately only if there are no interactions.

We stress again that the usefulness of GAMs is not confined to binary data. Indeed, they can handle any of the data types that GLMs are used for (e.g. Gaussian, multinomial, Poisson data) as well as certain types of survival data. The essential difference is simply that linear functions of the variables are replaced by unknown smooth functions which gives additional flexibility for the modelling process.

GAMs were proposed by Hastie & Tibshirani (1986) who implemented the method in the FORTRAN program GAIM (Generalized Additive Interactive Modelling). The program allows the replacing of any $f_j(x_j)$ by linear functions (i.e. the fitting of GLMs instead of GAMs), and enables one to test whether the unspecified smooth functions f_j can be replaced by one or more

parametric terms (see below). This allows an attractive approach to modelling, with a seamless transition between nonparametric and parametric models. Thus, a mix of parametric and nonparametric variables can be modelled together. App. 3 gives information on how GAIM may be obtained. GAMs will be also be able to be fitted in the next release of New S (Becker, Chambers & Wilks 1988). This version will be capable of using higher dimensional smoothers and the software will be described in Chambers & Hastie (in press).

GAMs require high speed computing and good graphics, and can be a powerful exploratory tool for the ecologist. For example, the detection of bimodality is difficult using GLMs but automatic for GAMs. Each variable is represented separately, therefore each function can be plotted separately to examine the roles of the variables in predicting the response. However, the effect of each variable is determined while allowing for the effects of the other variables (assuming there is no interaction). Complications caused by interactions will be discussed below.

Other forms of nonparametric curve fitting have been used in the ecological literature [e.g. the use of moving averages by Ashton (1976), Austin, Cunningham & Fleming (1984) and Ogden & Powell (1979), and running medians by Austin & Austin (1980)]. However GAMs have three advantages over such attempts. Firstly, it is not necessary to group continuous variables. Secondly, GAMs enable exploration of the simultaneous effect of several variables on $p(\boldsymbol{x})$ rather than just one. Lastly, GAMs allow hypothesis testing upon the variables. In the past, only limited use of smoothers were made in ecology. However, they are now becoming more widely used, especially since computers and software have become more available to biologists. They are a powerful tool, especially in an exploratory phase.

A number of techniques similar to GAMs have been developed relatively recently. These are briefly described in App. 1.

Technical details

In this section technical details for the development of a model using GAMs are described. For more details see Hastie & Tibshirani (1986, 1987a, 1987b, 1990).

Smoothers require the choice of a span in order to operate. The span measures how large a neighbourhood to take about a point when smoothing: the larger the span, the smoother the curve. It is chosen from the data. In all our examples the span was selected by cross validation, an automatic procedure which minimizes the sum of squared prediction errors when each data point is predicted using the rest of the data. More explicitly, the cross-validation sum of squares is minimized with respect to the span, w:

$$\mathrm{CVSS}(w) = (1/n)\sum_{i=1}^{n}\left(y_i - \hat{f}_w^{-i}(x_i)\right)^2$$

where $\hat{f}_w^{-i}(x_i)$ is the value of the smoother with span w at x_i obtained using all the data except the point (x_i, y_i). One can think of $\hat{f}_w^{-i}(x_i)$ as the estimate of y_i obtained using span w from the remainder of the data. The best span is the one that provides the 'best' estimates, in terms of the sum of squared prediction errors, on average. As each variable is smoothed separately, different variables may have different spans and smoothers. Another example of flexibility is to use a higher (but not too high) dimensional smoother to model complex interactions between variables (see below). In the case of two variables this would mean fitting $g(\mu) = \alpha + f(x_1, x_2)$.

The degrees of freedom of a smooth fit is usually a real number rather than an integer for a GLM. In the case of GAMs, the theory for this area has not been fully developed (see Buja, Hastie & Tibshirani 1989): but an approximation for running-lines smoothers is that $1 + \frac{1}{\mathrm{SPAN}} \leq$ degrees of freedom $\leq 2 + \frac{1}{\mathrm{SPAN}}$. For a parametric fit each term in the model takes up one degree of freedom.

After the fitting of a GAM, the nonlinearity of each covariate can be tested by fitting two models - one with a linear fit for the variable in question and the other with a nonparametric fit. The difference in deviance between the two models is attributed to nonlinearity. It is tested against a chi-square distribution with d degrees of freedom, where d is the difference in the degrees of freedom of the two models. Such tests are approximate and are conducted because parametric fits are preferred to nonparametric fits where possible (see below). This hypothesis testing is very similar to that of GLMs. An important special case is when there is only a single explanatory variable. In the case of binary data (y_i, x_i) (i = 1,.., n), a convenient method of testing whether the response curve is symmetric bell-shaped or not is to test for a significant difference between the GAM and Gaussian logit. This directly addresses a major problem of the past. In practice, several link functions should be tried instead of just one. This is because different link functions will give slightly different bell-shaped curves, which may make a difference. Common link functions for binary data include:

$g(p) =$	$\log\{p/(1-p)\}$	logit link
$g(p) =$	$F^{-1}(p)$	probit link
$g(p) =$	$\log\{-\log(1-p)\}$	complementary log-log link
$g(p) =$	$-\log\{-\log(p)\}$	log-log link,

where F is the distribution function of a standard normal

random variable. For all of these link functions, fitting $g(p) = \alpha + \beta_1 x + \beta_2 x^2$ will give a symmetric bell-shaped curve.

In software, the choice of smoother to implement is based on a number of criteria. The speed of a smoother is vitally important as smoothing is a major component of the computation. Theoretical properties related to the degrees of freedom and convergence is simplified if the smoother belongs to a class called symmetric shrinking smoothers. Also, a smoother that can handle observations with weights is mandatory. The version of GAIM used here can either use a local linear smoother or a cubic spline smoother. We used the former. In the future it is likely that software will contain a number of smoothers to choose from and that there will be documentation highlighting their differences.

An important point to note is that if a parametric curve is 'statistically allowable', then for reasons of parsimony it is to be preferred to a nonparametric curve. A drawback of nonparametric models is that important parameters such as the optimum and tolerance cannot in general be extracted from a nonparametric fit. Indeed, the concept of ecological amplitude becomes less well-defined for asymmetric curves. To remedy this apparent shortcoming of GAMs, it is necessary to fit a parametric model. Often, the plots of $\hat{f}_j(x_j)$ versus x_j suggest possible transformations of the variables. After transformation it may be possible to replace $\hat{f}_j(x_j)$ by a parametric curve. If this cannot be done for all the variables, then some variables can be fitted parametrically and the rest nonparametrically. Models such as this are called generalized partially additive models by McCullagh & Nelder (1989) and are examples of semi-additive models (Green & Yandell 1985; Stone 1986).

In the process of variable selection, GAMs have an advantage over GLMs in that the smoother automatically takes into account the shape of the curve for that variable. Thus, it is not necessary to choose whether a higher order term should be included, a decision that needs to be made for each case when using a GLM. For example, an x^2 term may be insufficient when attempting to add variable x into the model; including an x^3 term could have made x statistically significant. However, in the development of this work, we have encountered data sets, which when $g(\mu) = \alpha + f_1(x_1)$ was fitted, the function $\hat{f}_1(x_1)$ was $\hat{\beta}_{2\,1} x_1$ (i.e. the GAM was a GLM), but after fitting $g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_1^{\,2}$ it was discovered that β_2 was statistically significant. Such cases depend critically on the smoother used.

Approximate confidence curves can be obtained for each function of a fitted GAM. This is sometimes useful because it gives an indication which parts of the function are less accurately estimated (often because of fewer data points). GAIM can output confidence curves of the form $\hat{\boldsymbol{f}}_j \pm 2\mathrm{SD}[\hat{\boldsymbol{f}}_j]$ (see Fig. 3) and an example is given below. When fitting GLMs, partial residuals (Landwehr, Pregibon & Shoemaker 1984) can be used to identify the functional form of each variable (i.e. to identify the $f_j(x_j)$). In GAMs, partial residuals may also be calculated and plotting them can reveal outliers.

In order for a GAM to be successfully fitted to data, the algorithm requires convergence in two stages. The first is in the approximation of the regression surface by a sum of smooth functions. The second is in the overall fit of the model, as measured by the deviance. The former is called the backfitting algorithm (Friedman & Stuetzle 1981) and is imbedded within the the latter stage (the local scoring algorithm). Convergence is achieved by iterating until the changes are sufficiently small. Occasionally there is a failure to converge within a certain number of iterations and it seems more likely when the number of variables in the model is large. (Theoretical convergence properties have been derived for a particular class of smoothers called linear smoothers; Buja, Hastie & Tibshirani 1989).

A p-dimensional smoother can be used to model the regression surface (p throughout this paper denotes the number of explanatory variables). A p-dimensional smoother smooths a single response variable with respect to p other variables. However, this is not done very often - despite this being the most nonparametric way of proceeding. There are several reasons. Firstly, smoothers break down in high dimensions because of the "curse of dimensionality" (Friedman & Stuetzle 1981). Here, for fixed n, the data become more isolated in p-space and smoothers require a larger neighbourhood to find enough data points in order to calculate the variance of an estimate. Hence the estimate is no longer local and can be severely biased. Secondly, higher dimensional smoothers are numerically more intensive. Thirdly, it is difficult to interpret the effect of the variables on the response. The fitting of a set of additive one-dimensional smooths to approximate this surface is a compromise between a p-dimensional smoother and estimating the regression surface by the sum of linear functions.

An example of the use of a two-dimensional smoother is found in Huntley, Bartlein & Prentice (1989). In their study, the response variable was pollen percentage for *Fagus* and two explanatory variables were used (mean January and July temperatures). A two-dimensional smoother (LOESS; see Cleveland & Devlin 1988) was used to estimate the regression surface. Since the identity link function was used i.e. $g(\mu) = \mu$ the regression surface was the same as the response surface. In our notation they fitted $g(\mu) = \mu = \alpha + f(x_1,x_2)$ where Y = pollen abundance of *Fagus* (%) and $\boldsymbol{X}^T$=(mean January temperature, mean July temperature). Climate estimates were fitted and grey-scale plots similar to ours were also

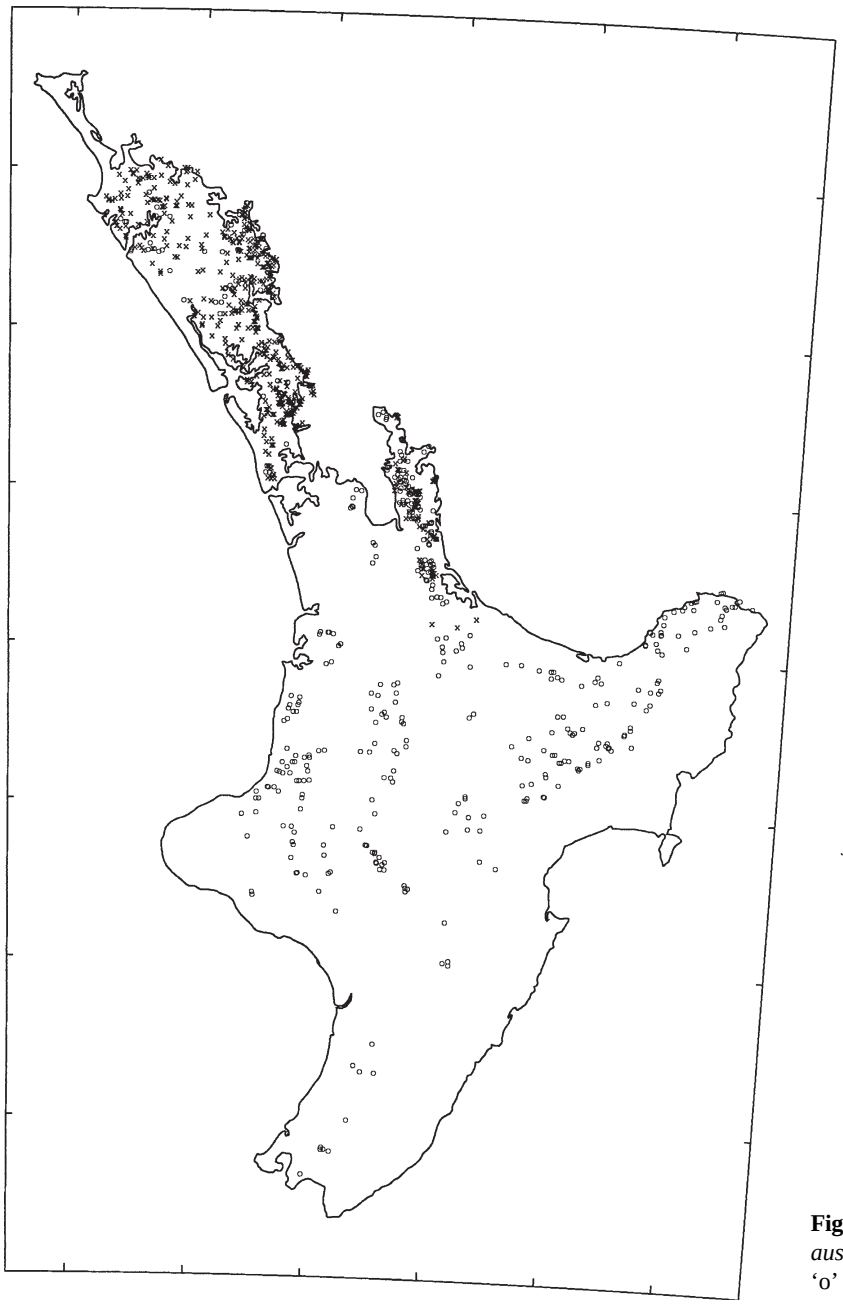

Fig. 1. Sampled site locations used in the *Agathis australis* study. An 'x' denotes a case site and a 'o' is a control site.

constructed.

In practice, it is often necessary to deal with interactions between variables. For example, if high rain compensates high temperatures in the survival of a certain plant species, then an interaction model would be required. Two variables do not interact if the effect on the response variable of one variable is the same no matter what value the second variable takes. Geometrically, cutting cross-sections of the regression surface at two points on the second axis and parallel to the first will yield parallel curves. When there is no interaction, the regression surface can be expressed as the sum of p functions $f_1(x_1)$, $f_2(x_2)$,..., $f_p(x_p)$. When interaction is present, the effect of a variable differs or depends in some way on the value of the other variable. In practice we try to detect and test for interactions and only include interactions in a model if the data reveals unequivocally that significant interactions are present. A common method of modelling simple interactions (for continuous variables) is to create a new variable which is the product of the two (i.e. if X_1 and X_2 are suspected to interact, the variable $X_3 = X_1X_2$ can be created and $g(\mu) =$

Table 1. Climate variables and abbreviations used in the study.

SOLAR RADIATION VARIABLES			
x_1	= srann	=	Annual mean solar radiation (MJ m^{-2} day^{-1})
x_2	= srmax	=	Highest monthly solar radiation
x_3	= srmin	=	Lowest monthly solar radiation
x_4	= srseas	=	Solar radiation seasonality (=$(x_2 - x_3)/x_1$)
x_5	= srwet	=	Mean solar radiation in the wettest quarter
x_6	= srdry	=	Mean solar radiation in the driest quarter
TEMPERATURE VARIABLES			
x_7	= tmann	=	Annual mean temperature (°C)
x_8	= tmmax	=	Maximum temperature of the hottest month
x_9	= tmmin	=	Minimum temperature of the coldest month
x_{10}	= tmseas	=	Temperature seasonality (=$(x_8 - x_9)/x_7$)
x_{11}	= tmwet	=	Mean temperature in the wettest quarter
x_{12}	= tmdry	=	Mean temperature in the driest quarter
RAINFALL VARIABLES			
x_{13}	= rfann	=	Annual precipition (mm)
x_{14}	= rfmax	=	Precipitation in the wettest month
x_{15}	= rfmin	=	Precipitation in the driest month
x_{16}	= rfseas	=	Precipitation seasonality (=$(x_{14} - x_{15})/(x_{13}/12)$)
x_{17}	= rfwet	=	Precipitation in the wettest quarter
x_{18}	= rfdry	=	Precipitation in the driest quarter.

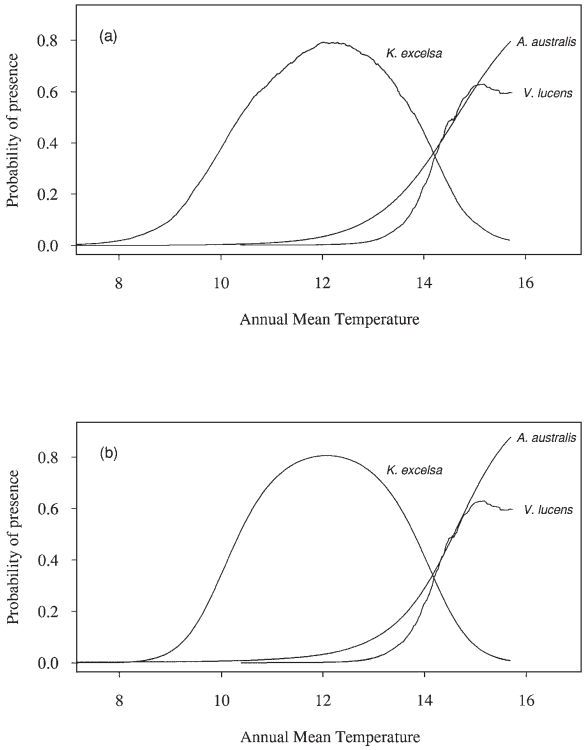

Fig. 2. Estimated presence-absence response curves for three contrasting species against annual mean temperature. (a) Results from a GAM with span chosen by cross-validation. (b) Results after tests for nonlinearity for *Knightia excelsa* and *Vitex lucens* in (a).

$\alpha + f_1(x_1) + f_2(x_2) + f_3(x_3)$ fitted. A test would then be made to see if X_3 is required).

In this study climate is assumed to be the major determinant in species' distributions; however, qualitative variables such as soil type may also be important and can be used as explanatory variables. For example, suppose there were I soil types and one continuous variable, such as annual mean temperature. It would be possible to fit logit(p) =$\alpha_i + f_i$ (annual mean temperature) (i = 1,.., I). This assumes that the effect of annual mean temperature depends on the soil type. A simpler model would be to fit logit(p) = $\alpha_i + f$(annual mean temperature) (i = 1,.., I) where the effect of annual mean temperature is the same, no matter what the soil type. These models parallel the analysis of covariance in linear regression theory. Hastie & Tibshirani (1986, 1987a, 1987b, 1990) give examples illustrating how to detect interactions.

Data sources

Plant species used

Knightia excelsa is a common, long-lived Angiosperm tree, found throughout the North Island of New Zealand and the north of the South Island. It is found from sea level to 1050 m, growing on a wide range of substrates and landforms. It is considered to be a hardy species and may be found as an early emergent during succession, as well as a canopy member of closed forest.

Agathis australis is a common, very long-lived Gymnosperm tree, restricted to the area north of 38°S and West of 176°E. It is found from sea level (although usually away from the coast) to an elevation of 700 m, typically growing in shallow, nutrient-poor soils on steep slopes. *Agathis australis* is considered to be moderately frost sensitive and its geographic distribution is assumed to reflect this sensitivity.

Vitex lucens is a common Angiosperm tree with a similar geographic distribution to *Agathis australis*, except that it occurs across the full width of the North Island. It has a more coastal/lowland distribution, being found from exposed coasts to an elevation of 450m. *Vitex lucens* may be found on a variety of slopes, although it typically occurs on more moderate slopes that have higher nutrient status soils than for *Agathis australis*. It is considered to be extremely frost tender, with a more common occurrence on north-facing slopes.

All three species may be found occurring at the same sites, although there are clearly local niche differences as well as a more regional scale differentiation.

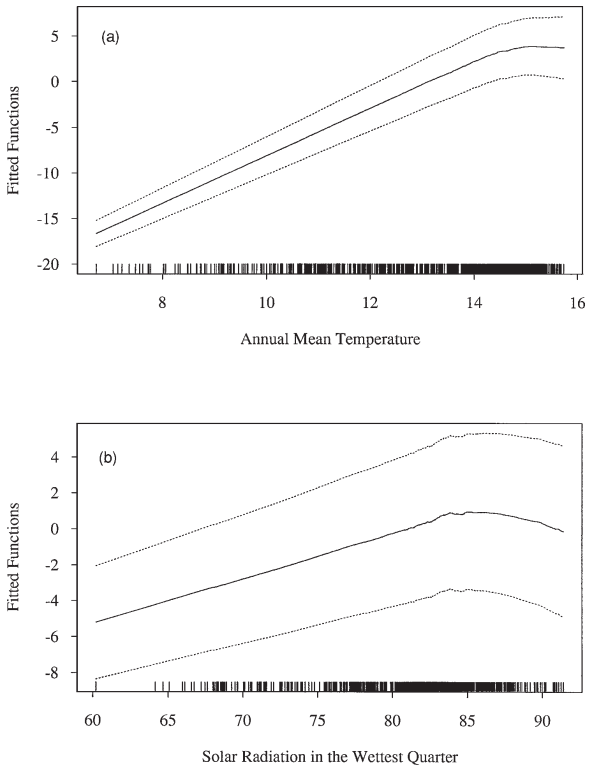

Fig. 3. (a) The estimated contribution $\hat{f}_7(x_7)$ of annual mean temperature to logit$\{p(\boldsymbol{x})\}$ for *Vitex lucens* is the solid curve. The dotted curves are $\hat{f}_7(x_7) \pm 2\mathrm{SD}[\hat{f}_7(x_7)]$. Each vertical bar at the base of the graph denotes an observation with that value. (b) The estimated contribution $\hat{f}_5(x_5)$ of solar radiation in the wettest quarter to logit$\{p(\boldsymbol{x})\}$ for *Vitex lucens* is the solid curve. The dotted curves are $\hat{f}_5(x_5) \pm 2\mathrm{SD}[\hat{f}_5(x_5)]$. Each vertical bar at the base of the graph denotes an observation with that value.

Species data sources

The data sets were derived by merging data from herbaria and a number of vegetation surveys to document the distribution of established trees. A proportion of the data (approximately 60% for *Knightia excelsa*, 30% for *Agathis australis* and 10% for *Vitex lucens*) was collected during the 1950's by the New Zealand Forest Service (Masters, Holloway & McKelvey 1957) as part of a stratified survey of timber resources. The rest of the data were either collected during comprehensive surveys from 1979-1986 (the Northland Forest Inventory of the then Department of Lands and Survey, and the Protected Natural Area programme of the now Department of Conservation), or represent additional observations by one of us (NDM). The data were collected in a variety of ways ranging from plot counts through cover abundance of variable areas to presence/absence. In total, 1464 records were usable for *Agathis australis*, 4172 for *Knightia excelsa* and 1239 for *Vitex lucens*. A consequence of the variable origin of the data was that geocoded presence/absence information proved to be the only reliable basis for further analysis.

The spatial intensity of the sample sites varies, partly as a consequence of forest clearances, but also in part, as a consequence of survey design (Fig. 1). However, in total, these data cover the full topographical and geographical range of all species and should encompass the full environmental variation within which the species survive. Although not a perfect random sample of sites, the sites are believed to be a good representation of these species in New Zealand.

Because the entire data set was too large to be accommodated, a case-control sample (see App. 2) was taken, making the analysis slightly more complicated.

Climate data

Climate variables were estimated for each location by use of BIOCLIM (Nix 1986), a program which estimates parameters from surfaces generated using a Laplacian smoothing spline fitted to meteorological data (Hutchinson & Bischof 1983; Hutchinson et al. 1984). Estimation of climate variables at particular sites from fitted climate surfaces were developed for New Zealand by Mitchell (1991) and are functions of latitude, longitude and elevation. From each of the temperature, rainfall and solar radiation surfaces, the following six variables were produced: the annual mean or total, highest mean monthly maximum, lowest mean monthly minimum, seasonality ((monthly maximum - monthly minimum)/annual value), mean or total for the driest and wettest quarters. The variables are listed in Table 1. For some of the variables, notably precipitation, the correlation between variables is quite high.

This type of data has been used elsewhere to analyse species distributions with respect to climate (e.g. Nix 1986; Busby 1986; Caughley et al. 1987; Podger et al. 1990; Mitchell 1989, 1991). However, in most of these cases no attempt was made to statistically define which climatic variables were of greatest importance, nor to spatially relate the probability of survival to specific climatic effects.

Analyses and Results

GAMs are illustrated using presence-absence data of several North Island species. Different facets of GAMs are illustrated by three separate analyses. The first is a GAM fitted for three species against a single gradient.

Table 2. (a) GAM fitted on *Vitex lucens*. (b) Gaussian logit model with an interaction term fitted on *Vitex lucens*. S.E. is the standard error of the estimate and Z is the test statistic value.

(a)

Variable	Span	d.f.	Slope
constant	-	1	0.5671
tmann	0.7	2.334	smooth
srwet	0.7	2.373	smooth

Deviance = 457.909; Degrees of freedom = 821.29.

(b)

Variable		Estimate	S.E.	Z
constant	α	−224.550	118.954	−1.89
tmann	β_1	6.3261	8.908	0.71
(tmann)^2	β_2	−0.7890	0.2264	−3.48
srwet	β_3	3.8333	2.026	1.89
(srwet)^2	β_4	−0.04027	0.01143	−3.52
(tmann)(srwet)	β_5	0.21575	0.08500	2.54

Deviance = 456.363; Degrees of freedom = 821.

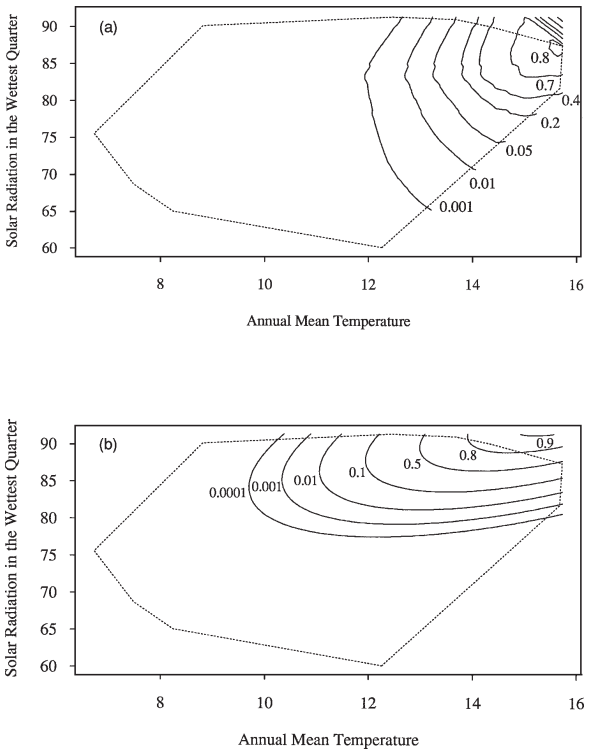

Fig. 4. (a) Contour plot of the response surface (probability of presence) from a GAM fitted on *Vitex lucens*. Units for the variables given in Table 1. The dotted polygon represents a convex hull surrounding the sample data.
(b) Contour plot of the response surface (probability of presence) from a Gaussian logit model with an interaction term, (eqn. (4)), fitted on *Vitex lucens*. Units for the variables given in Table 1. The dotted polygon represents a convex hull surrounding the data. The contour levels match those in (a).

The second fits a species against two gradients (explanatory variables) and interactions are explored. Both of these will be compared to their GLM equivalent. The third analysis is a larger analysis where all statistically significant variables are added to form a model and interactions tested. A grey-scale plot is constructed and some applications are briefly mentioned. It is important to appreciate that the one and two variable models are included only to illustrate what is happening during the initial stages of a full model-building analysis and to illustrate concepts in a simple setting.

Three contrasting species

Fig. 2a shows the response curves of GAMs fitted for the three contrasting species against annual mean temperature. This variable was chosen because temperature variables, as compared to the solar radiation or rainfall variables, were always the most statistically significant single variables in the regressions.

Two important points are illustrated by Fig. 2. Firstly, a GAM may be a GLM if the smoother fits a linear function (smoothers tend to fit lines to data if the data is linear or if it is very noisy). This is shown in the case of *Agathis australis* in Fig 2a. Hence the nonparametric fit is, in fact, a parametric one. Where, as here, only part of the curve is visible over the existing range of climatic conditions, the curve is described as being truncated.

The second point is an attempt to test nonlinearity of the covariate. This can only be done with the other two species. *Knightia excelsa* has a classical bell-shape curve which suggests that the Gaussian logit model (2) may be appropriate. After fitting (2) and observing the reduction in deviance, it was concluded that the parametric fit is indeed 'allowable'. The same was done for *Vitex lucens* and the nonparametric fit could not be rejected: *Vitex lucens* has to be fitted nonparametrically. The final curves for the three species are found in Fig. 2b.

Vitex lucens

Several methods of variable selection were available. One alternative is 'subsets selection' (see Miller 1984, 1990). In this case, a number of models containing one, two, three, and so on, variables are examined which are considered the 'best' according to some specified criterion. Secondly, since several variables were highly correlated, particularly those within the same

group (temperature, solar radiation and precipitation), it would be possible to use one representative from each group.

Stepwise regression (see Seber 1977) was adopted here as a third alternative. The procedure was started with a constant term: at each stage the most significant variable (if any) is added and a test is made as to whether any variables in the model can be deleted. This is a combination of forward selection and backward elimination. Stepwise regression has a number of pitfalls (see Greenland 1989) and its use in this paper is not intended as an unconditional recommendation.

The most significant variable is not necessarily the variable that causes the greatest drop in deviance since variables may have different degrees of freedom. Instead, the drop in deviance is compared to a chi-square distribution with degrees of freedom equal to the degrees of freedom between the two models. For example, a drop of six in deviance with an extra one degree of freedom is statistically more significant than a drop of 10 in deviance with three degrees of freedom ($0.014 = P(\chi_1^2 > 6) < P(\chi_3^2 > 10) = 0.019$). As the degrees of freedom for GAMs is real-valued, the ability to compute areas under a χ^2 probability density function with real degrees of freedom is therefore required. If the drop in deviance was less than a suitable cut-off such as the 5% point of a χ^2 distribution (with degrees of freedom equal to the degrees of freedom of that variable), then that variable could not be added to the model. Such a test is valid if the sample size is large. In this study a 5% cut-off was used throughout.

The mean annual temperature was found to be the best fitting single variable, giving a large D^2 value of 57.4%. D^2 (= 100 (null deviance – deviance)/null deviance)) represents the percentage of deviance explained and is analogous to R^2 in regression. This was followed by solar radiation in the wettest quarter (Table 2a). Confidence curves as described above, with the fitted function, are plotted in Fig. 3. These results suggest that the overall temperature environment is of major importance for the survival of the species. Simply, the warmer the location, the more suitable it is for the species. The levelling off of the fitted temperature function at 15 °C suggests this to be its optimum, as was also seen in Fig. 2. From what little is known of the species, this tends to conform to the observation that it is commoner on warm, north-facing slopes. The role of solar radiation in the wettest quarter suggests that the species must maintain a minimum level of metabolism during the winter. This can only be met in specific localities, although the sites of highest winter solar radiation appear to be less suitable.

The interaction between the variables was tested firstly by adding the variable (tmann)(srwet) to the model. This resulted in a drop in deviance of 3.533 for an extra 1.016 degree of freedom, giving some evidence of an interaction (p = 0.062). The response surface resulting from this model is shown as a contour plot (Fig. 4a). Overlaid on it is a convex hull (Green 1981), approximately indicating the region in which the sample data lie.

Table 3. Stages of the stepwise regression procedure in fitting the model for *Agathis australis*. Note: x_9 = minimum monthly temperature, x_3 = minimum monthly solar radiation, x_{18} = rainfall in the driest quarter, x_6 = solar radiation in the driest quarter.

Variable	Deviance	d.f.	D^2
intercept	1243.682	901	
intercept+x_9	703.568	900	.4343
intercept+x_9+x_3	635.128	897.65	.4893
intercept+x_9+x_3+x_{18}	591.008	894.28	.5248
intercept+x_9+x_3+x_{18}+x_6	575.262	892.44	.5375

As with the one variable case, the (no-interaction) GAM was compared to its equivalent GLM. The response surface was modelled by a two-dimensional Gaussian logit model and in this case provided as good a fit as the GAM model, provided an interaction term was included. The results of this regression,

$$\text{logit}(p) = \alpha + \beta_1 x_7 + \beta_2 x^2_7 + \beta_3 x_5 + \beta_4 x^2_5 + \beta_5 x_7 x_5, \quad (4)$$

are found in Table 2b and Fig. 4b. The Z column provides an approximate method for testing for the significance for that variable (i.e. whether the coefficient is zero). It is obtained by dividing the estimate by its standard error. This test-statistic is standard normal if the coefficient is zero. The results show that the interaction term is statistically significant (p = 0.0111). This p-value is smaller than before, reflecting the increased power of a parametric test, and to some extent the approximate nature of inference for GAMs.

From Table 2b, the optimum (u_1, u_2) can be estimated (Jongman, ter Braak & van Tongeren 1987):

$$\hat{d} = 4\hat{\beta}_2\hat{\beta}_4 - \hat{\beta}_5 = 0.0805,$$

$$\hat{u}_1 = \left(\hat{\beta}_5\hat{\beta}_3 - 2\hat{\beta}_1\hat{\beta}_4\right)\Big/\hat{d} = 16.59 \ (^\circ\text{C})$$

$$\hat{u}_2 = \left(\hat{\beta}_5\hat{\beta}_1 - 2\hat{\beta}_3\hat{\beta}_2\right)\Big/\hat{d} = 92.05 \ \left(\text{MJm}^{-2}\text{day}^{-1}\right)$$

Note that this point lies outside the convex hull of the data points and so should be treated with caution. A biological interpretation might be that there are presently no sites in our sample which provide an optimum environment for this species. In comparison ($\hat{u}_1$, $\hat{u}_2$) is quite different from the optimum suggested by the GAM

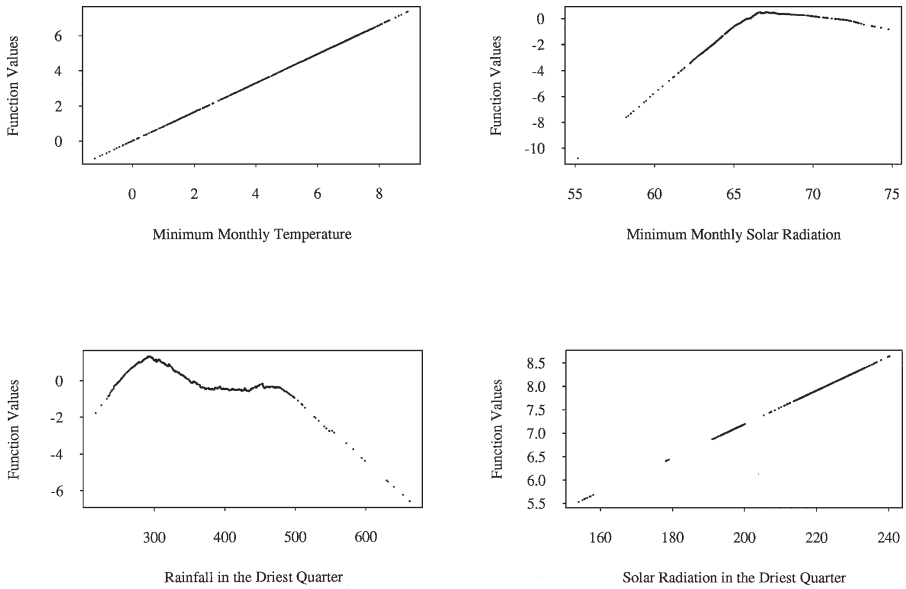

Fig. 5. Fitted function values, $\hat{f}_j(x_j)$, of each observation on *Agathis australis* vs. each variable x_j (j = 3,6,9,18). Confidence curves have been omitted.

(Fig. 4a). It is well known that it is difficult to estimate the optimum accurately if it lies 'outside' of the data (ter Braak & Looman 1986). The ability to calculate an explicit optimum clearly illustrates why a parametric fit is preferred to a nonparametric one.

Because the 'interpretation' of the fitted functions $\hat{f}_j(x_j)$ only makes sense when there are no interactions, we have assumed there are no interactions. Unfortunately, the two-dimensional Gaussian logit model provides strong evidence of an interaction - thus suggesting our interpretation of each climate variable may not be accurate. Nevertheless, insight has been gained, especially from an exploratory point of view. If the analysis was terminated at this point, (4) would probably be satisfactory for most purposes.

Agathis australis

From the initial 18 variables, the final model for *Agathis australis* was obtained by stepwise regression. The final model and the path leading to it is shown in Table 3. An analysis of partial residuals showed that there were a number of outliers in the data. These observations were deleted early in the analysis.

Four variables were found to provide the best fit: minimum monthly mean temperature and solar radiation, total rainfall of the driest quarter and mean solar radiation in the driest quarter. Two of these were linear functions on the logit scale (minimum monthly mean temperature and solar radiation in the driest quarter) and the other two were approximately unimodal. The fitted function values are plotted in Fig. 5, where the confidence curves have been omitted. Interactions were ex-

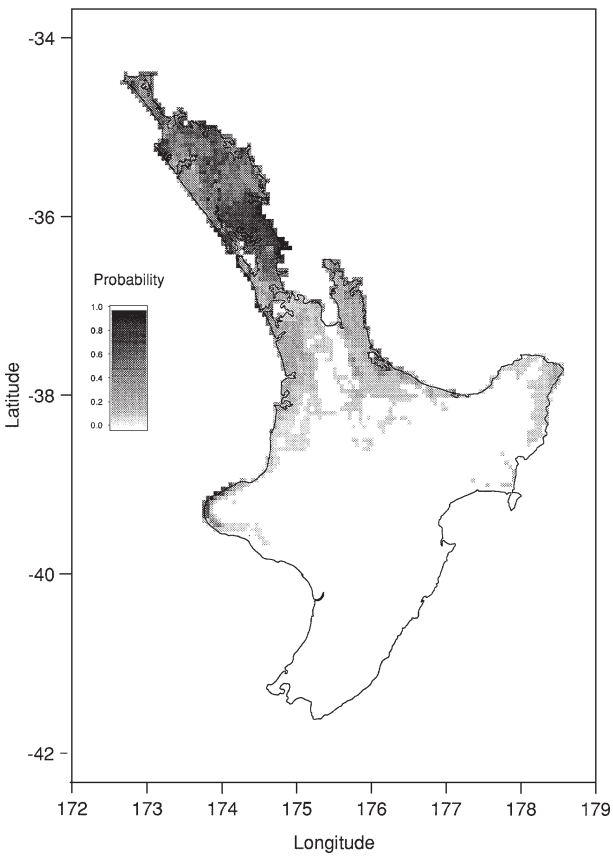

Fig. 6. Grey-scale plot of the fitted GAM (found in Table 3) on *Agathis australis* overlaid upon a map of the North Island of New Zealand. Areas where the fitted probability is less than 0.01 are white.

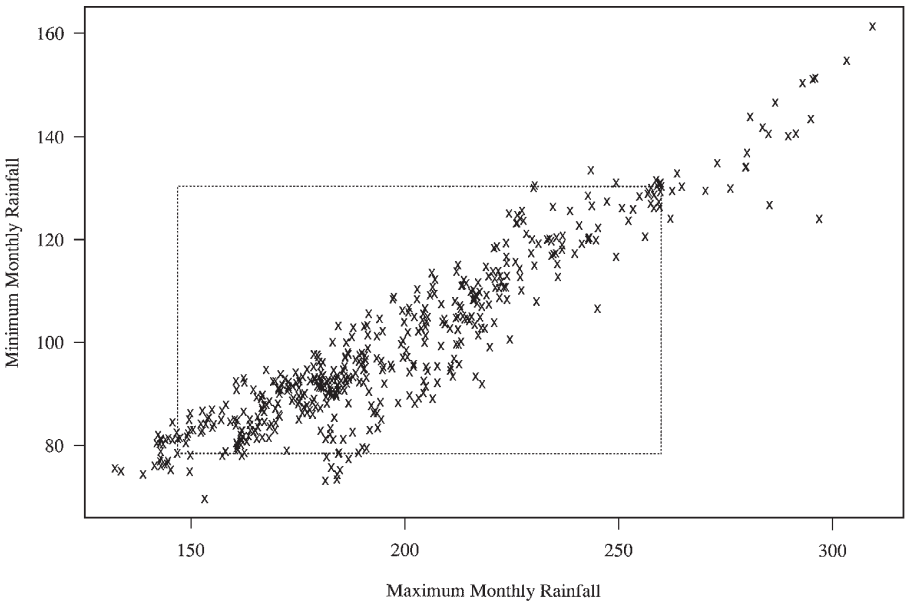

Fig. 7. Case sites used in the *Agathis australis* study. The endpoints of the box for each variable are the 5 and 95 percentiles e.g. the 5 percentile of maximum monthly rainfall is 150 mm. Correlation coefficient = 0.933.

plored by separately testing the significance of the six created variables resulting from each pairwise product of the four variables. None of them needed to be added; so it was concluded that these four variables were sufficient to adequately model the data on *Agathis australis*.

The nature of these variables and the order in which they were added to the model are of great interest. The first two variables relate to mid-winter conditions. The positive slope for minimum monthly temperature suggest the importance of mid-winter temperatures in controlling the species' distribution. It was interesting that minimum solar radiation was the second variable to be added. This suggests that, as for *Vitex lucens*, in mid-winter the plants are not 'dormant' and that active metabolism must proceed. Again, the positive slope suggests a limiting value, below which the plants cannot survive. However, the fitted function also suggests that at higher values of mid-winter solar radiation, a site will become less suitable for survival. The biological significance of this is not clear.

The other two variables relate to the driest quarter, which is usually from November to January - the main growing season. The unimodal nature of the rainfall curve suggests that too much rain is as bad as too little rain during this period. The latter effect is not unexpected, since other work suggests that summer drought halts wood production (Palmer 1982). The possible effect of too much rainfall is more problematic. Higher rainfall areas are principally at higher altitude and it is possibly related to cooler growing conditions, or even reduced solar radiation due to increased cloudiness. There may also be biological effects such as incidence of disease or increased competition. The plateau for rainfall in the driest quarter is probably due to the sampling. The solar radiation response suggests the species requires a 'high light' environment during the growing season, presumably to maintain growth and reproduction. Work by Bieleski (1959) and Pook (1979) have shown that *Agathis australis* seedlings grow best under higher levels of illumination. The common occurrence of the species on exposed ridges (especially in the southern part of its range) would also tend to suggest that the species can make use of these high light environments.

Once the final model was obtained, a probability map was constructed for the the North Island. To achieve this, the fitted climatic surfaces were interpolated on a five km^2 topographic grid of the North Island using BIOCLIM and the variables were substituted into the final model. As some of the functions were smooths, numerical interpolation was used. The probability surface can be displayed by either a contour plot or a grey-scale plot. The latter is more intuitive, while the former makes it easier to interpolate a value at a specific location. The grey-scale plot is shown in Fig. 6. It suggests the most suitable environment for *Agathis australis* is found midway in Northland, i.e. at approximately 36°S. Interestingly, locations to the south of its present range (ca. 39°S; near New Plymouth) and further east than at present (ca. 178°E; in the Bay of Plenty-East Cape

region) also appear to be suitable.

The idea of plotting probabilities over a map is not new: Wrigley (1977a,b) produced these and called them probability surface maps. We believe that the approach used here is a great improvement over these earlier attempts, both methodologically and graphically.

Discussion

GAMs are a major improvement over previously described methods for analysis of the relationship between species distribution and climate. Previously, Busby (1986), Booth (1985) and Mitchell (1991) have all used the same 'BIOCLIM approach', and although the techniques showed some success, the methods employed were somewhat arbitrary, and consisted of fitting a *p*-dimensional box over the data - of which some variables are highly correlated. Even in two dimensions, this can be seen as unsuitable (Fig. 7). In these studies there were no measures of the magnitude of suitability of a location, simply that it had potential to support the species.

Work by Booth et al. (1987) employed a Gower similarity measure

$$d_{ij} = \sum_{k=1}^{S} |X_{ik} - X_{jk}|$$

to assess the similarity in climate between two locations. However, their method did not adjust for the magnitude of the variables and their results were based upon a few of the precipitation variables. We believe that even if a more appropriate similarity measure is used, GAMs or GLMs are more elegant and flexible.

There are several unresolvable problems with the analysis illustrated in this paper. Firstly, we never know the exact values of the climatic variables at each site - they were only estimated. Standard regression theory is based on the assumption that all explanatory variables are measured without error (Chatterjee & Hadi 1988); this assumption also holds for GAMs. This may have introduced an additional source of error and bias in our study, resulting in the attenuation of regression effects. Secondly, spatial autocorrelation violates the independence assumption required by GLMs (and GAMs). In the notation used earlier, the Y_i are not independent: the presence/absence of a species may be affected by its presence/absence in nearby locations. Thirdly, a random sample of all possible sites (which excludes farmland, etc.) is required which, for a study on as a large scale as this, becomes almost impossible to obtain. In Fig. 1, which shows the sites used in the *Agathis australis* analysis, the non-randomness is apparent. For example, there is a disproportionate number of sites in the Coromandel Peninsular (37°S and 175.5°E in Fig. 6). One way to improve the results is to perform a stratified case-control study. The country would be stratified into regions and a test can be made for a 'region effect'. Theory for this method, for GLMs, is found in Scott & Wild (1989); how this theory might change with GAMs has not yet been clarified. Fourthly, the size of the sites were not identical; indeed, some site sizes were unknown. It was only known that the species were present at a specific location (the accuracy of location was always to approximately 100m, which on a regional scale such as this, effectively makes each record a point location). Known site sizes varied from 20 m^2 to about 50 m^2.

The method by which the site locations are chosen is crucial. In our opinion many of the different types of response curves seen in the literature could be due to the sampling (see Austin, Cunningham & Fleming 1984). Indeed, by oversampling and undersampling certain areas, any type of response curve could theoretically be obtained. Under random sampling, all forested areas of the country should have an equal probability of being sampled. The samples would be all the same size and recording the presence-absence of each species should be made without any classification errors (Ekholm & Palmgren 1982).

Agathis australis, although a very interesting species, was not necessarily a good species to study because its distribution has been severely affected by humans. Its advantage is that the present day distribution is very well documented and although its total abundance has been severely reduced (Beever 1981), it still occurs throughout its full range. We also chose this species so that the improvement obtained by using GAMs could be compared with an earlier analysis by Mitchell (1991).

There are many potentially important applications of GAMs. Clearly, the grey-scale plots indicate into which areas a species could be introduced. Also, GAMS provide a refined approach to modelling the potential changes in species distributions that might result from global warming. The approach would be to fit GAMs separately to each species and substitute the climate estimates of the area into the fitted models to estimate the change in the probability of presence for each species. This approach has been attempted previously (e.g. work by J. R. Leathwick in Hollinger 1990 using GLMs for an area in the central North Island of New Zealand). Another application would be to fit GAMS to several species to see which species would be the most (climatically) suitable to introduce into a particular area (e.g. Booth 1985, 1990). An agricultural application would be to grow crops at various experimental sites reflecting a wide range of environments, with their productivity

measured on an ordinal scale (for example, 0 for non-survival and 10 for abundant growth). Since GAIM allows the fitting of proportional odds models (Hastie & Tibshirani 1987b), finer estimates of crop performances with respect to the climatic variables could be determined. As a result, this could be used to find regions climatically suitable for use in species introduction.

Acknowledgements. We thank Drs. T. J. Hastie and R. J. Tibshirani for providing us with GAIM and Dr. C. J. Wild for many valuable suggestions. We also thank John Nicholls and John Leathwick of the Forest Research Institute, Rotorua, for access to the National Forest Survey records. NDM would like to thank the Auckland University Research Committee for providing funds associated with this work and the NZ Lotteries Board for travel funding. We also thank the three anonymous referees.

References

Anon. 1987. *The Generalized Linear Interactive Modelling System: The GLIM system.* Release 3.77. Royal Statistical Society, London.

Anon. 1988. *SAS Guide for Personal Computers*, Version 6.03. SAS Institute Inc., Cary, NC.

Ashton, D. H. 1976. The vegetation of Mount Piper, Central Victoria: a study of a continuum. *J. Ecol.* 64: 463-483.

Austin, M. P. 1979. Current approaches to the non-linearity problem in vegetation analysis. In: Patil, G. P. & Rosenzweig, M. (eds.) *Satellite Program in Statistical Ecology. S12 Contemporary Quantitive Ecology and Related Ecometrics*, pp. 197-210, International Co-operative Publishing House, Fairland, Maryland.

Austin, M. P. 1987. Models for the analysis of species' response to environmental gradients. *Vegetatio* 69: 35-45.

Austin, M. P. & Austin, B. O. 1980. Behaviour of experimental plant communities along a nutrient gradient. *J. Ecol.* 68: 891-918.

Austin, M. P. & Cunningham R. B. 1981. Observational analysis of environmental gradients. *Proc. Ecol. Soc. Aust.* 11: 109-119.

Austin, M. P., Cunningham R. B. & Fleming, P. M. 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio* 55: 11-27.

Austin, M. P., Nicholls, A. O. & Margules, C. R. 1990. Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. *Ecol. Monogr.* 60: 161-177.

Austin, M. P. & Smith, T. M. 1989. A new model for the continuum concept. *Vegetatio* 83: 35-47.

Becker, R. A., Chambers, J. M. & Wilks, A. R. 1988. *The New S Language.* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California.

Becker, R. A., Chambers, J. M. & Wilks, A. R. 1990. *The New S Language: a Programming Environment for Data Analysis and Graphics.* Statistical Sciences, Inc., Seattle, Washington.

Beever, J. 1981. A map of the pre-European vegetation of lower Northland, New Zealand. *N.Z J. Bot.* 19: 105-110.

Bieleski, R. L. 1959. Factors affecting the growth and distribution of Kauri (*Agathis australis* Salisb.) III. Effect of temperature and soil conditions. *Aust. J. Bot.* 7: 279-294.

Booth, T. H. 1985. A new method for assisting species selection. *Commonw. For. Rev.* 64: 241-250.

Booth, T. H. 1990. Mapping regions climatically suitable for particular tree species at the global scale. *For. Ecol. Manage.* 36: 47-60.

Booth, T. H., Nix, H. A., Hutchinson, M. F. & Busby J. R. 1987. Grid matching: a new method for homoclime analysis. *Agric. For. Meteorol.* 39: 241-255.

Breiman, L. & Friedman, J. H. 1985. Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Assoc.* 80: 580-597.

Buja, A., Hastie, T. & Tibshirani, R. 1989. Linear smoothers and additive models. *Ann. Statist.* 17: 453-555.

Busby, J. R. 1986. A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Austr. J. Ecol.* 11:1-7.

Caughley, G., Short, J., Grigg, G. C. & Nix, H. A. 1987. Kangaroos and climate: an analysis in distribution. *J. Animal Ecol.* 56: 751-761.

Chambers, J. M. & Hastie, T. In press. *Statistical Models in S.* Wadsworth, Pacific Grove, CA.

Chatterjee, S. & Hadi, A. S. 1988. *Sensitivity Analysis in Linear Regression.* John Wiley & Sons, New York.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* 74: 829-836.

Cleveland, W. S. & Devlin, S. J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Assoc.* 83: 596-610.

Dixon, W. J. 1987. *BMDP Statistical Software.* University of California Press, Berkeley.

Dobson, A. J. 1990. *An Introduction to Generalized Linear Models.* Chapman and Hall, London.

Ekholm, A. & Palmgren, J. 1982. A model for a binary response with misclassifications. In: Gilchrist, R. (ed.) *GLIM82: Proceedings of the International Conference on Generalised Linear Models*, pp. 128-143, Lecture Notes in Statistics, 14. Springer-Verlag, New York.

Friedman, J. H. & Stuetzle, W. 1981. Projection pursuit regression. *J. Am. Statist. Assoc.* 76: 817-823.

Goodall, C. 1990. A survey of smoothing techniques. In: Fox, J., Long, J. S. (eds.) *Modern Methods of Data Analysis*, Sage Publications, Newbury Park, CA.

Green, P. J. 1981. Peeling bivariate data. In: Barnett, V. (ed.), *Interpreting Multivariate Data*, pp. 3-20, John Wiley & Sons, New York.

Green, P. J. & Yandell, B. 1985. Semi-parametric generalized linear models. In: Gilchrist, R., Francis, B., Whittaker, J. (eds.) *Proceedings of the GLIM 1985 Conference*, pp. 44-55, Lecture Notes in Statistics, 32. Springer-Verlag, Berlin.

Greenland, S. 1989. Modelling variable selection in epidemiologic analysis. *Am. J. of Public Health* 79: 340-349.

Hastie, T. & Tibshirani, R. 1986. Generalized additive models

(with discussion). *Statist. Sci.* 1: 297-318.
Hastie, T. & Tibshirani, R. 1987a. Generalized additive models: some applications. *J. Am. Statist. Assoc.* 82: 371-386.
Hastie, T. & Tibshirani, R. 1987b. Non-parametric logistic and proportional odds regression. *Appl. Statist.* 36: 260-276.
Hastie, T. J. & Tibshirani, R. J. 1990. *Generalized Additive Models.* Chapman and Hall, London.
Hollinger, D. Y. 1990. Forestry and forest ecosystems. In: *Climatic Change: Impacts on New Zealand*. pp. 66-77. Ministry for the Environment, Wellington.
Hosmer, D. W. & Lemeshow, S. 1989. *Applied Logistic Regression*. John Wiley & Sons, New York.
Huntley, B., Bartlein, P. J. & Prentice, I. C. 1989. Climatic control of the distribution and abundance of beech (*Fagus* L.) in Europe and North America. *J. Biogeogr*. 16: 551-560.
Hutchinson, M. F. & Bischof, R. J. 1983. A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the Hunter Valley, New South Wales. *Aust. Meteorol. Mag.* 31: 179-184.
Hutchinson, M. F., Booth, T. H., McMahon, J. P. & Nix, H. A. 1984. Estimating monthly mean values of daily total solar radiation for Australia. *Solar Energy* 32: 277-290.
Jongman, R. H. G., ter Braak, C. J. F. & van Tongeren, O. F. R. 1987. *Data Analysis in Community and Landscape Ecology*. Pudoc, Wageningen.
Landwehr, J. M., Pregibon, D. & Shoemaker, A. C. 1984. Graphical methods for assessing logistic regression models (with discussion). *J. Am. Statist. Assoc.* 79: 61-81.
Masters, S. E., Holloway, J. T. & McKelvey, J. 1957. *The National Forest Survey of New Zealand*, 1955. New Zealand Forest Service, Wellington.
McCullagh, P. & Nelder, J. A. 1989. *Generalized Linear Models*, 2nd ed. Chapman and Hall. London.
Miller, A. J. 1984. Selection of subsets of regression variables (with discussion). *J. R. Statist. Soc. A.* 147: 389-425.
Miller, A. J. 1990. *Subset Selection in Regression.* Chapman and Hall. London.
Mitchell, N. D. 1989. The analysis of species distributions with respect to climate: past, present and future. In: Sansom, J. (ed.) *Proceedings of the Fourth International Meeting on Statistical Climatology*, pp. 211-214, N. Z. Meteorological Service, Wellington.
Mitchell, N. D. 1991. The derivation of climate surfaces for New Zealand, and their application to the bioclimatic analysis of the distribution of Kauri (*Agathis australis*). *N. Z. J. Royal Society* 21: 13-24.
Nelder, J. A. & Wedderburn, R. W. M. 1972. Generalized linear models. *J. R. Statist. Soc.* A 135: 370-84.
Nicholls, A. O. 1989. How to make biological surveys go further with generalized linear models. *Biol. Conserv.* 50: 51-75.
Nix, H. 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore, R. (ed.) *Atlas of Australian Elapid Snakes*, pp. 4-15, Bureau of Flora and Fauna, Canberra: Aust. Govt. Publishing Service, Canberra.
Ogden, J. & Powell, J. A. 1979. A quantitive description of forest vegetation on an altitudinal gradient in the Mount Field National Park, Tasmania, and a discussion of its history and dynamics. *Aust. J. Ecol.* 4: 193-325.
Palmer, J. 1982. *A dendrochronological study of Kauri (Agathis australis).* Unpubl. MSc Thesis, University of Auckland.
Payne, R. W., Lane, P. W. et al. 1987. *Genstat 5 Reference Manual*, Oxford University Press, New York.
Podger, F. D., Mummery, D. C., Palzer, C. R. & Brown, M. J. 1990. Bioclimatic analysis of the distribution of damage of native plants in Tasmania by *Phytophthora cinnamomi*. *Aust. J. Ecol.* 15: 281-289.
Pook, E. W. 1979. Seedling growth in Tanekaha (*Phyllocladus trichomanoides*): effects of shade and other seedling species. *N. Z. J. For. Sci.* 9: 193-200.
Prentice, R. L. & Pyke, R. A. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66: 403-411.
Scott, A. J. & Wild, C. J. 1986. Fitting logistic models under case-control or choice based sampling. *J. R. Statist. Soc. B*, 48: 170-182.
Scott, A. J. & Wild, C. J. 1989. Hypothesis testing in case-control studies. *Biometrika* 76: 806-808.
Scott, A. J. & Wild, C. J. 1991. Fitting logistic regression models in stratified case-control studies. *Biometrics* 47: 497-510.
Seber, G. A. F. 1977. *Linear Regression Analysis.* Wiley, New York.
Stone, C. J. 1986. The dimensionality reduction principle for generalized additive models. *Ann. Statist.* 14: 590-606.
ter Braak, C. J. F. & Gremmen, N. J. M. 1987. Ecological amplitudes of plant species and internal consistency of Ellenberg's indicator values for moisture. *Vegetatio* 69: 79-87.
ter Braak, C. J. F. & Looman, C. W. N. 1986. Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio* 65: 3-11.
van Houwelingen, J. C. & le Cessie, S. 1988. Logistic regression, a review. *Statistica Neerlandica* 42: 215-232.
Whittaker, R. H. 1956. Vegetation of the Great Smoky Mountains. *Ecol. Monogr.* 26: 1-80.
Wilkinson, L. 1987. *SYSTAT: The System for Statistics.* Systat Inc., Evanston, IL.
Wrigley, N. 1977a. Probability surface mapping: a new approach to trend surface mapping. *Trans. Inst. Brit. Geogr.* 2: 129-140.
Wrigley, N. 1977b. *Probability Surface Mapping: An Introduction with examples and Fortran Programs. Concepts and Techniques in Modern Geography, No. 16*, Geo. Abstracts Ltd, Norwich.

Received 11 August 1990;
Revision received 15 May 1991;
Accepted 20 June 1991.

App. 1. This appendix briefly describes three modern regression methods somewhat similar to GAMs. A general reference is Becker, Chambers & Wilks (1990).

Recall that for GLMs

$$g\{E(Y|\boldsymbol{x})\} = \alpha + \beta^T\boldsymbol{x} = \alpha + \sum_{j=1}^{p} \beta_j x_j$$

and for GAMs

$$g\{E(Y|\boldsymbol{x})\} = \alpha + \sum_{j=1}^{p} f_j(x)_j$$

where g is prespecified by the user. GAMs are therefore a way of extending the additive model

$$E(Y|\boldsymbol{x}) = \sum_{j=1}^{p} f_j(x)_j.$$

Alternating Conditional Expectation (ACE; see Breiman & Friedman 1985) provides another way of extending the additive model. For ACE one fits the model

$$E\{h(Y)|\boldsymbol{x}\} = \sum_{j=1}^{p} f_j(x)_j. \qquad (5)$$

where h, f_1, f_2,.., f_p are smooth nonlinear functions estimated from the data ($\boldsymbol{y}$, $\boldsymbol{x}_1$,.., $\boldsymbol{x}_p$). The criterion is to maximize the correlation between both sides of (5). As with GAMs it is very useful as a modelling tool to help determine which of the response y and explanatory variables x_1, x_2,.., x_p are in need of a nonlinear transformation and what type of transformation is needed. A method similar to ACE is additive nonlinear regression with variance stabilisation (AVAS). It also fits a model of the form (5) with the f_j chosen as in ACE, but the h chosen to achieve constant residual variance. It has the same usefulness as ACE.

Another extension of the additive model above is to fit

$$E(Y|\boldsymbol{x}) = \sum_{k=1}^{K} f_k(a_k^T\boldsymbol{x}),$$

where for each k = 1,.., K, $\boldsymbol{a}_k$ is a unit vector, f_k is a smooth nonlinear function and $a_k^T\boldsymbol{x}$ is the projection of $\boldsymbol{x}$ (= (x_1, x_2,.., x_p)^T) onto a_k. The a_k are found by a numerical search and K is determined from the data. This is Projection Pursuit Regression (PPR; Friedman & Stuetzle 1981) and, like ACE and AVAS, is available in the S-PLUS package. Although PPR is difficult to interpret, it easily handles interactions. FORTRAN source code for these three methods is available from the statlib library (see App. 3).

The theory for these methods has not been developed to the extent of GLMs. As a result its current use is principally for exploratory data analysis and with formal analyses usually made under the framework of GLMs. All three are sensitive to outliers and hence lack the robustness often necessary in biological analyses.

App. 2. Case-control (or retrospective) studies are popular in biostatistics. The essential idea is that separate samples of cases and controls are taken. In this field, a case is often a person with a certain disease and a control is someone without the disease. In our context we will define a case (respectively, control) to be a site with the species present (respectively absent). Sampling was necessary to keep the computation time down to within reasonable levels. A case-control sample was performed to maximize information about the effect of the variables.

To illustrate the necessity of a case-control study, suppose there were 600 sites where a species existed and 9400 sites where it did not. If 1000 sites were sampled (unconditionally) from all the sites, there would be approximately 60 case sites and 940 control sites. This provides much less information about the effect of the covariates than separate samples of 500 case sites and 500 control sites.

For logit models (i.e. $g(p)$ = logit(p) in (1)), one can ignore the fact that case-control sampling has been undertaken except for a simple correction to the intercept term in the regression. To obtain the correct intercept term, $\log(n_1N_0/(n_0N_1))$ is subtracted from the fitted intercept term , where n_0=number of controls in the subsample, N_0 = number of controls in the whole data set, n_1 = number of cases in the subsample and N_1=number of cases in the whole data set (see Scott & Wild 1991). This adjustment is necessary in order to correctly scale the probabilities due to the different sampling intensities of the cases and controls.

In our analyses, we took random samples of about 450 cases and 450 controls for each species. For example, for *Agathis australis*, n_0 = 419, N_0 = 7139, n_1 = 493 and N_1 = 1464. For further information about case-control sampling see Scott & Wild 1986, Prentice & Pyke 1979, and Hosmer & Lemeshow 1989.

App. 3. The version of GAIM used in this paper was the original 1984 version. Currently, two later versions are available. One is PC GAIM, a version for PC's running DOS. Copies may be ordered by writing to S. N. Tibshirani Enterprises, Inc. 5334 Yonge St, Suite 1714. Toronto, ON M2N 6M2, Canada.

The FORTRAN source code of a second version of GAIM is available from the statlib library. This is free, and can be obtained by sending an e-mail letter to *STATLIB@-LIB.STAT.CMU.EDU* consisting of a single line "send gamfit from general". A list of other statistical software available can be obtained by sending the line "send index from general" to the same address, or the line "send index" for more general information.